DOCUMENT RESUME

ED 233 056 TM 830 503

AUTHOR Phillips, S. E.
TITLE Logistic Achievement Test Scaling and Equating with Fixed versus Estimated Lower Asymptotes.
PUB DATE Apr 83
NOTE 17p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Montreal, Quebec, April 12-14, 1983).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Comparative Analysis; *Equated Scores; Estimation (Mathematics); Latent Trait Theory; *Maximum Likelihood Statistics; *Measurement Techniques; Psychometrics; *Scaling; Standardized Tests
IDENTIFIERS Equipercentile Equating; *LOGIST Computer Program; Lower Asymptotes; Test Batteries; *Vertical Equating

ABSTRACT

This study compared the lower asymptotes estimated by the maximum likelihood procedures of the LOGIST computer program with those obtained via application of the Norton methodology. The study also compared the equating results from the three-parameter logistic model with those obtained from the equipercentile, Rasch, and conditional three-parameter models for vertically equating test levels when the scaling method was applied to a multilevel achievement test battery. Two grades, fourth and eighth, and two subtests, reading and mathematics, were selected for analysis. For each content area, approximately 500 students per grade completed both the scaling test and the regular battery test. There appeared to be no evidence to recommend the three-parameter logistic model over the two-parameter modified logistic model when the scaling test approach was used to equate multilevel achievement batteries. Each method appeared to provide a different definition of score-to-equivalence. The moderate rank correlations between lower asymptote estimates for those few items for which estimates were possible in the three-parameter model suggested that the Norton methodology may provide a reasonable alternative for estimating lower asymptotes when sample sizes are small (PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED233056

LOGISTIC ACHIEVEMENT TEST SCALING AND EQUATING WITH FIXED VERSUS ESTIMATED LOWER ASYMPTOTES

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

S. E. Phillips
Michigan State University

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

S. E. Phillips

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the NCME Annual Meeting
Montreal, 1983

Printed in the U.S.A.

TM 830 503

LOGISTIC ACHIEVEMENT TEST SCALING AND EQUATING WITH FIXED VERSUS ESTIMATED LOWER ASYMPTOTES

One of the purposes of standardized achievement test batteries is to measure growth across grade levels. To do so, test publishers construct a common score scale to which the regular battery test forms at each grade level are equated. Typically, this vertical equating process involves common items administered to adjacent grade levels (anchor test method) or the administration of a sample of content from a wide range of grade levels to students at every grade (scaling test method). Mittman (1958) found the scaling test method most satisfactory for scaling achievement tests and some major test publishers rely on it in whole or in part (Hieronymus, Lindquist, and Hoover, 1978; Cole, Trent, and Wadel, 1982).

Traditionally, publishers have used equipercentile equating methods to link adjacent test levels or to link scaling and regular battery tests. More recently, item response theory methods have been used for this purpose. However, the evidence on application of item response theory models to multi-level achievement test batteries is conflicting and incomplete.

In considering the anchor test approach to the equating issue, Rentz and Bashaw (1977), Wright and Stone (1978), Beard and Pettie (1979), and Guskey (1979) reported favorable results for the Rasch model. Andersen (1973), Slinde and Linn (1978), Loyd and Hoover (1980), and Divgi (1981) presented results which conflicted with the other studies. Hambleton and Traub (1973) and Hoover (1982) preferred the results obtained with a two-parameter logistic model. Marco, Petersen, and Stewart (1979), Lord (1980), Kolen (1981), and Cook and Douglass (1982) recommended the three-parameter

logistic model. Waller (1980) reported that the three-parameter model is more expensive to implement in test development work.

In a previous study (Phillips, 1982), the author compared the equipercentile Rasch, one-parameter modified and two-parameter logistic models for vertically equating test levels when the scaling test method was applied to a multi-level achievement test battery. Two grades, fourth and eighth, and two subtests, reading and mathematics, were selected for analysis. The two-parameter item response theory model as modified in the study provided results consistent with the equipercentile method.

The modification referred to in the previously referenced study involved the specification of a conditional three-parameter logistic model in which the lower asymptotes of the item characteristic curves were not estimated by a maximum likelihood procedure. The lower asymptotes in the model were fixed based on a method and logic developed by Norton (1950). Norton demonstrated that the proportion of students choosing the least attractive distractor for an item is an upper bound for the proportion of students answering the item correctly due to random guessing. Standard item analysis data from the norming sample, which is routinely generated by a test publisher, provided the Norton proportions used to estimate the lower asymptotes of the item characteristic curves.

The present study is an extension of this work. Its purpose was two-fold. First, the study compared the lower asymptotes estimated by the maximum likelihood procedures of the LOGIST computer program (Wood, Wingersky, and Lord, 1976) with those obtained via application of the Norton methodology. Second, the study compared the equating results from the three-parameter logistic model with those obtained from the equipercentile, Rasch, and conditional three-parameter models of the original study (1982).

In the original study, the full three-parameter model was not included in part due to cost and in part due to the small sample sizes involved. Lord (1980) suggests sample sizes of at least two thousand for adequate estimation of the lower asymptotes of the item characteristic curves in a three-parameter model. The equating comparisons in the study were based on final samples of 300-500 valid records of students who had taken both the scaling and regular battery tests. The relatively small sample sizes occurred because use of the scaling test method requires a special test administration apart from the national norming of the test. Sampling and cost considerations often prohibit publishers from using larger samples in such a study.

Despite the fact that "good" results could not be expected from the three-parameter model due to the small sample sizes, it was still of interest to determine whether any of the lower asymptotes could be estimated from the available data, the degree to which the maximum likelihood estimates were similar to the Norton methodology estimates, and whether any improved conguence to equipercentile equating results could be obtained using the three-parameter model. Even if none of the lower asymptotes could be estimated from the data and were instead set to a constant based on the number of item distractors, there was no apriori evidence to suggest that this procedure would provide any "worse" equating for this particular application. The lower asymptotes are fixed values in both procedures; the only differences being the methods and assumptions by which the estimation occurred.

Methodology

The procedures of the author's original study (1982) were followed in the present study so that the results obtained would be comparable to those

of the original study. The three parameter logistic model was applied to test data from the same four subject/grade samples.

The test utilized in the study is an achievement test battery designed for use in grades K through 12. There are eleven different levels of the test; Levels 6 through 14 correspond to grades K through 8, Level 15/16 to grades 9 and 10, and Level 17/18 to grades 11 and 12.

The reading, language, and mathematics subtests consist of 15, 35, and 40 items, respectively, in grades 3-12. These grades are allowed 40, 30, and 28 minutes for completing each subtest and are required to record their answers on a separate machine-scorable answer sheet.

The main standardization of the achievement test was completed in the spring of 1980. Approximately 8,000 students per grade in grades K-12 were tested during the month of April.

Participating school districts were randomly sampled from listings compiled from census data based on geographic location and district size. Both public and Catholic districts were included in the sample. Within each school district, buildings were randomly sampled. In general, all students attending a sampled school building were tested in the standardization.

The scaling test in each content area (reading and mathematics) consisted of a sample of representative content across grade levels. The mathematics scaling test consisted of 68 items and the reading scaling test consisted of 63 four-choice multiple-choice items. Both were administered in a fifty-minute time period. All of the scaling test items were different from all items in the regular battery tests.

Preceding the spring standardization of the regular battery tests, a subsample of participating schools was selected to administer the scaling test forms. For each content area, approximately 500 students per grade completed both the scaling test and the regular battery test. The same

reading and mathematics scaling test forms were given to both fourth and eighth grade students.

Unlike other tests and studies of this type, there were no items in common between levels of the achievement test battery or between scaling test forms and their corresponding achievement subtests. Thus, once the grade development scale (GDS) had been constructed from the scaling test data, it was only possible to link the grade development scale to achievement test scores via the group of students at each grade who had taken both tests.

In the original study, the maximum likelihood estimates of the item difficulty and discrimination parameters were obtained using the LOGIST (Wood, Wingersky, and Lord, 1976) computer program. When LOGIST is used to estimate a three-parameter model, c values are set to the reciprocal of the number of choices per item minus .05 if $b - 2/a < -2.0$ (c= lower asymptote, b = difficulty, a = discrimination). Warm (1978) estimates that up to eighty percent of test items may fall in this category.

Recent communications with Marilyn Wingersky at Educational Testing Service (1982) and Ronald Hambleton at the University of Massachusettes (1982) indicate that improved estimation of the lower asymptotes of the item characteristic curves is possible with the new version of LOGIST (Wingersky, Barton and Lord, 1982). However, since the new version of LOGIST was unavailable to the author at the time the study was completed, only the results from the old version of LOGIST have been reported. Results from the new version would be expected to differ significantly from those presented here.

The equating of tests in the original study and for the three-parameter model in the present study was completed using the estimated true score equating methodology outlined by Lord (1980). This procedure is based on the assumption of equal ability distributions and can estimate equated scores only as small as the sum of the lower asymptotes for the given test. An auxillary procedure for

equating scores below the chance level has been suggested by Lord (1980) but was not utilized in the present study. A more detailed description of the equating procedures is provided in the original study (Phillips, 1982).

To determine whether equating differences between any two methods were greater than expected by chance, it was necessary to estimate the amount of error associated with each method. The method error for the three-parameter model was estimated by conducting a reliability study based on the experimental design used in the original study. In that study, the regular battery test was equated to itself with random samples of 500 students. The degree to which any raw score in one sample was not equated to the same raw score in the other sample provided an estimate of the errors associated with the equating method.

For comparison on a common metric, all equating differences were expressed in terms of grade development scaled scores (GDS). Any other standard score scale would have been equally suitable for comparison purposes. This particular scale was chosen because it was available from the publisher and because many researchers have an intuitive idea of the meaning and instructional significance of grade equivalent score differences.

Results and Discussion

Frequency distributions of the lower asymptotes estimated by the Norton methodology for the two-parameter modified logistic model and by the LOGIST program for the three-parameter logistic model are presented in Tables 1A and 1B. Results for the mathematics scaling and regular battery tests are presented in Table 1A; lower asymptotes for the reading scaling and regular battery tests are listed in Table 1B. At the bottom of each table, the number of items, means, and standard deviations are given. The percent of estimable lower asymptotes for the three-parameter logistic model is also given in the summary statistics section at the bottom of each table.

Insert Tables 1A and 1B about here

As expected, the majority of items had unestimable lower asymptotes for the three-parameter logistic model. The percent of items having lower asymptotes set at .20 was greatest for the regular battery tests and varied from .0 to 22 percent. The majority of items with estimable lower asymptotes were fourth grade scaling test items.

A comparison of the lower asymptotes by method for items having an estimable lower asymptote in the three-parameter logistic model is presented in Table 2. There were a total of 26 items with estimable lower asymptotes in the three-parameter logistic model across subjects, grades, and test types. of those items 23 were scaling test items. The classical difficulty of each item is reported in the last column of Table 2 and indicates that all the items with estimable lower asymptotes in the three-parameter model had difficulty values of .30 or less. The Spearman rank correlation between the two-parameter modified and three-parameter estimates was .374 which was significantly greater than zero at $p < .03$. If the rank correlation between the two- and three-parameter lower asymptote estimates is recomputed omitting three-parameter estimates of .18 and .19 (values very close to the 0.20 at which the other item lower asymptotes were set), the result is .586 which is also significant at $p < .03$. Though few items had estimable lower asymptotes in the three-parameter logistic model, the three-parameter estimates for those that did appear to have some similarity to the lower asymptotes estimated via the Norton methodology in the two-parameter modified logistic model.

Insert Table 2 about here

Equating results for the three-parameter model (LOG3P) are presented in Table 3 by subject and grade. The results for the equipercentile (EQUIPER), Rasch, logistic one-parameter modified (LOG1PM), and logistic two-parameter modified (LOG2PM) methods of the original study (Phillips, 1982) are also included in Table 3 for comparison. The reported statistic is the mean absolute difference between GDS (grade development scaled score) values assigned by each equating method across the raw score range of each test. The numbers in parentheses along the diagonals in Table 3 are the error estimates obtained in the reliability study.

Insert Table 3 about here

Mean absolute differences from the experimental studies which were considered "significant" are followed by an asterisk in Table 3. "Significant" in this context was defined as a mean absolute difference which exceeded the error estimate for both of the equating methods being compared. For example, the entry .0889 representing the mean absolute difference in GDS values assigned by the equipercentile and three-parameter logistic methods is asterisked since it is greater than both the equipercentile (.0094) and three-parameter (.0148) error estimates.

For three out of the four subject/grade combinations, the three-parameter logistic method provided results which were significantly different from the equipercentile results. In all but one case, the magnitude of the average absolute difference from the equipercentile method was slightly smaller for the three-parameter than for the two-parameter modified logistic method. In two out of the four subject/grade combinations, the three-parameter results differed significantly from those of the two-parameter modified logistic method. As in the results of the original study, the largest differences occurred in the Rasch comparisons in reading grade 4.

In summary, there appears to be no evidence to recommend the three-parameter logistic model over the two-parameter modified logistic model when the scaling test approach is used to equate multilevel achievement batteries. Each method appears to provide a different definition of score-to-score equivalence. Further study of the relationship between instructional level and the GDS scores assigned by each method is needed to establish which scale provides the "better" definition of appropriate instructional level. Knowing that each method represents a unique definition of score equivalence provides no guidelines for judging which definition best serves test users in practical applications.

The moderate rank correlations between lower asymptote estimates for those few items for which estimates were possible in the three parameter model suggest that the Norton methodology may provide a reasonable alternative for estimating lower asymptotes when sample sizes are small. Further study is needed to determine whether use of the new version of LOGIST can improve upon these results. With larger sample sizes or other types of achievement test data, the results may differ from those presented here.

It should also be noted that the error data for this study represents only one possible method of quantifying within method chance variation. Other error estimates may result in different conclusions.

TABLE 1A

Frequency Distributions of Estimated Lower Asymptotes

For The Two-Parameter Modified and Three-Parameter

Logistic Models

	Math Grade 4				Math Grade 8			
Lower	Regular		Scaling		Regular		Scaling	
Asymptote	LOG2PM	LOG3P	LOG2PM	LOG3P	LOG2PM	LOG3P	LOG2PM	LOG3P
.01	1		10				34	
.02	4		4		2		6	
.03	6		7		4		11	
.04	3		5		3		5	
.05	4		5	1	6		4	
.06	3		5	1	5		2	
.07	4		9		2		2	
.08	1		6	1	3			
.09	1		2		4			
.10	2		1	1	1			
.11	2		5	1	2		2	
.12	2		2	2			1	
.13	2		2	1			1	
.14			2		1			
.15			1		1			
.16			2		1			
.17				1				
.18				4				
.19		1		2				
.20		34		53		35		68
N	35	35	68	68	35	35	68	68
$\bar{X}$	.061	.200	.064	.186	.068	.200	.028	.200
SD	.035	.002	.041	.035	.035	.000	.028	.000
%estimatable	--	3	--	22	--	0	--	0

TABLE 1B

Frequency Distributions of Estimated Lower Asymptotes

For The Two-Parameter Modified and Three-Parameter

Logistic Models

	Reading Grade 4				Reading Grade 8			
Lower	Regular		Scaling		Regular		Scaling	
Asymptote	LOG2PM	LOG3P	LOG2PM	LOG3P	LOG2PM	LOG3P	LOG2PM	LOG3P
.01	2		9		2		20	
.02	8		4		7		5	
.03	5		1		7		5	
.04	10		3		4		6	
.05	9		5		9		3	
.06	3		3		4		7	
.07	4		6		6		4	
.08	3		6		1		3	
.09			2		4		5	
.10	2		4		1		1	
.11	1		2	1	2			
.12			2	1	1			
.13	2		3		1		2	
.14			2				2	
.15			5		1			
.16			1					
.17			4					
.18	1	1		1				
.19				4		1		1
.20		49		56		49		62
.21			1					
N	50	50	63	63	50	50	63	63
$\bar{X}$	.053	.200	.082	.193	.056	.200	.046	.200
SD	.033	.003	.052	.007	.032	.001	.036	.001
%estimable	--	2	--	11	--	2	--	2

TABLE 2

Comparison of Lower Asymptotes by Method for Items Having An Estimable Lower Asymptote in the Three-Parameter Logistic Model

			LOG2PM	LOG3P	P-Value
Math	Grade 4	Regular	.11	.19	.26
		Scaling	.05	.05	.05
			.05	.06	.05
			.07	.08	.08
			.09	.10	.09
			.10	.11	.10
			.08	.12	.11
			.12	.12	.12
			.07	.13	.14
			.16	.17	.16
			.07	.18	.22
			.03	.18	.22
			.11	.18	.18
			.06	.18	.19
			.08	.19	.25
			.07	.19	.21
Math	Grade 8	Regular	---	---	---
		Scaling	---	---	---
Reading	Grade 4	Regular	.18	.18	.30
		Scaling	.15	.19	.25
			.10	.19	.25
			.07	.12	.15
			.21	.18	.21
			.11	.11	.11
			.14	.19	.20
			.14	.19	.19
Reading	Grade 8	Regular	.08	.19	.30
		Scaling	.14	.19	.24

Table 3

MEAN ABSOLUTE DIFFERENCES IN GDS UNITS BY METHOD FOR THE RELIABILITY AND EXPERIMENTAL STUDIES

		EQUIPER	RASCH	LOG1PM	LOG2PM	LOG3P
GRADE 4 MATH	EQUIPER	(.0094)	.1100*	.1167*	.1000*	.0889*
	RASCH		(.0188)	.1433*	.1333*	.1333*
	LOG1PM			(.0000)	.0300	.0333*
	LOG2PM				(.0375)	.0148
	LOG3P					(.0148)
GRADE 4 READ	EQUIPER	(.0404)	.2911*	.0956*	.1156*	.0615*
	RASCH		(.0000)	.2644*	.2778*	.2769*
	LOG1PM			(.0489)	.0244	.0333
	LOG2PM				(.0340)	.0436*
	LOG3P					(.0154)
GRADE 8 MATH	EQUIPER	(.0719)	.0813*	.1031*	.0688	.0852*
	RASCH		(.0000)	.0594*	.0688*	.0926*
	LOG1PM			(.0000)	.0781*	.1111*
	LOG2PM				(.0313)	.0704*
	LOG3P					(.0667)
GRADE 8 READ	EQUIPER	(.0766)	.1000*	.0745	.0532	.0462
	RASCH		(.0979)	.1021	.1106*	.1103*
	LOG1PM			(.1106)	.0426	.0641
	LOG2PM				(.1085)	.0487
	LOG3P					(.0949)

* "significant" = greater than reliability values for both methods being compared

REFERENCES

Andersen, E.B. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 129-140.

Beard, J.G. & Pettie, A.L. A comparison of linear and Rasch equating methods results for basic skills assessment tests. Paper presented to the American Educational Research Association Annual Meeting, San Francisco,1979.

Cole, N.S., Trent, E.R. & Wadel D.C. The 3-R's Test Technical Manual. Achievement Edition. (Grades K-12). Chicago, Illinois, The Riverside Publishing Company, 1982.

Cook, L.L. and Douglass, J.B. Analysis of fit and vertical equating with the three-parameter model. Paper presented to the American Educational Research Association Annual Meeting, New York, 1982.

Divgi, D.R. Does the Rasch model really work? Not if you look closely. Paper presented to the American Education Research Association Annual Meeting, Los Angeles, 1981.

Guskey, T.R. Comparison of a Rasch model scale and the grade equivalent scale for the vertical equating of test scores. Applied Psychological Measurement, Vol. V, No. 2, Spring, 1981.

Hambleton, R.K. Personal Communication. University of Massachusettes, January, 1983.

Hambleton, R.K. and Traub, R.E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, Vol. 26, 1973.

Hoover, H.D. Personal Communication. University of Iowa, January, 1983.

Hieronymus, A.N., Lindquist, E.F., and Hoover, H.D. Technical Manual Iowa Tests of Basic Skills (Forms 7/8). Boston, Houghton Mifflin, 1978.

Kolen, M.J. Comparison of traditional and latent trait theory methods for equating tests. Journal of Educational Measurement. Vol. 18, 1981.

Lord, R.M. Application of Item Response Theory to Practical Testing Problems. Hillsdale, New Jersey, Lawrence Erlbaum Associates, 1980.

Loyd, B.H. and Hoover, H.D. Vertical equating using the Rasch model. Paper presented to the American Educational Research Association Annual Meeting, Boston, 1980.

Marco, G.L., Petersen, N.S., and Stewart, E.E. Applicability of two logistic models for equating test scores when tests and samples are varied. Paper presented to the American Educational Research Association Annual Meeting, San Francisco, 1979.

Mittman, A. An Empirical Study of Methods of Scaling Achievement Tests at the Elementary Grade Level. Unpublished doctoral dissertation, University of Iowa, Iowa City, Iowa, 1958.

Norton, D. W. A Suggested Method for Estimating the Proportion of a Group Selecting the Correct Response to a Multiple-choice Item by Guessing. Unpublished masters thesis, University of Iowa, Iowa City, Iowa, 1950.

Phillips, S. E. Comparison of Equipercentile and Latent Trait Equating When the Scaling Test Method is Applied to a Multilevel Achievement Battery. 1982 (in press).

Rentz, R. R. and Bashaw, W. L. The National Reference Scale for Reading: An Application of the Rasch Model. Journal of Educational Measurement, 1977, 14, 161-180.

Slinde, J. A. and Linn, R. L. An Exploration of the Adequacy of the Rasch Model for the Problem of Vertical Equating. Journal of Educational Measurement, 1978.

Waller, M. I. An Objective Procedure for Comparing the One, Two, and Three-Parameter Logistic Latent Trait Models. Paper presented to the American Educational Research Association Annual Meeting, Boston, 1980.

Wingersky, M. S. Personal Communication. Educational Testing Service, January, 1983.

Wingersky, M.S., Barton, M.A., and Lord, F.M. LOGIST User's Guide. Educational Testing Service, Princeton, N.J. 1982.

Wood, R.L., Wingersky, M.S., and Lord, F.M. LOGIST: A Computer Program for Estimating Examinee Ability and Item Characteristic Curve Parameters. Research Memorandum 76-6. Educational Testing Service, Princeton, N.J., 1976.

Wright, B.D. and Stone, M.H. Best Test Design: A Handbook for Rasch Measurement. Palo Alto: Scientific Press, 1978.